Empathy predicts false belief reasoning ability: Evidence from the N400

Heather J Ferguson

James E Cane

Michelle Douchkov

Daniel Wright

University of Kent

Correspondence to:
Heather Ferguson
School of Psychology
Keynes College
University of Kent
Canterbury
Kent CT2 7NP
England, UK

email: h.ferguson@kent.ac.uk
Tel: +44 (0) 1227 827120 Fax: +44 (0) 1227 827030

Word count: 5590 (excluding references, tables and figure legends)

Abstract

Interpreting others' actions relies on an understanding of their current mental state. Emerging research has begun to identify a number of factors that give rise to individual differences in this ability. We report an ERP study where participants (N=28) read contexts that described a character having a true (TB) or false belief (FB) about an object's location. A second sentence described where that character would look for the object. Critically, this sentence included a sentence-final noun that was either consistent or inconsistent with the character's belief. Participants also completed the Empathy Quotient questionnaire. Analysis of the N400 revealed that when the character held a TB about the object's location, the N400 waveform was more negative-going for belief inconsistent *vs.* belief consistent critical words. However, when the character held a FB about the object's location the opposite pattern was found. Intriguingly, correlations between the N400 inconsistency effect and individuals' empathy scores showed a significant correlation for FB but not TB. This suggests that people who are high in empathy can successfully interpret events according to the character's FB, while low empathisers bias their interpretation of events to their own egocentric view.

Key words: Theory of Mind, false beliefs, event-related brain potentials, N400, discourse comprehension

Introduction

Understanding others' beliefs, desires and intentions is a vital part of successful everyday social interaction (termed Theory of Mind, ToM). Typically, children develop these important social skills between the ages of 2 and 7 years old (Wellman et al., 2001). Failures to infer others' mental states are attributed to severe interference from one's own knowledge of reality, and difficulties inhibiting the egocentric perspective (de Villiers & Pyers, 2002; Flavell et al.,1990; Wellman & Bartsch, 1988). A common paradigm to assess ToM is the false belief task (Baron-Cohen et al., 1985), where participants are introduced to two characters: Sally and Anne. Sally puts a marble into her basket then goes out for a walk. In Sally's absence Anne takes the marble out of the basket, and puts it into her box. Participants must then answer test questions that require them to either infer Sally's false belief ("where will Sally look for the marble?") or to recall the narrative reality ("where is the marble really?"). This task requires individuals to see things from someone else's point of view (known as 'perspective-taking'), and relies heavily on ToM abilities to understand other peoples' mental states (which might be different from one's own), and how this might affect the other person's knowledge, beliefs and actions. In contrast to children, adults do not typically make errors on this task when traditional response-based measures are employed (i.e. question-answer), implying that they do not suffer interference from their own knowledge of reality. However, when more sensitive measures are used (e.g. reaction times, eye-tracking and brain responses), even healthy adults experience difficulties in considering other peoples' perspectives (Mitchell et al., 1996; Birch & Bloom, 2007). The current study examines the electrophysiological basis of the ability to understand events according to others' (false) beliefs, and explores for the first time how this is modulated by individual differences in social skills, namely the ability to empathise with others. Empathy is a multidimentional term, with some researchers conceptualising empathy as an affective/emotional response to another's mental state (e.g. Stotland, 1969), while others have viewed it in terms of the cognitive

mechanisms that enable us to understand others' perspectives (Dymond, 1949). The cognitive conceptualisation of empathy therefore overlaps with ToM, in that considering other peoples' minds is central to both and is therefore likely to influence false belief reasoning in the current study. Contemporary measures of empathy incorporate constructs from both of these dimensions, treating them as distinct but related subscales, thereby providing an overarching and integrated approach to empathy (see Baron-Cohen & Wheelwright, 2004; Davis, 1983).

Traditionally, research on ToM has examined the underlying processes at group-level, assuming that healthy adults perform in similar ways on these tasks. However, emerging research has begun to identify a number of key factors that give rise to individual differences in ToM ability, including mood (Converse et al., 2008), social relationships (Savitsky et al., 2011), cultural background (Wu & Keysar, 2007), autistic traits (Brunye et al., 2012; Kessler & Wang, 2012), and executive function skills such as working memory and inhibitory control (e.g. Brown-Schmidt, 2009; Cane et al., under review; German & Hehman, 2006; Lin et al., 2010). Though no studies to date have explicitly examined how individual differences in empathy might predict one's ability to interpret events according to others' (false) beliefs, existing evidence provides a number of reasons to suggest that such a relationship might exist. As described above, empathy is conceptually very similar to ToM with a related affective dimension, and has been described as an emotion-specific mentalising ToM ability (Tager-Flusberg & Sullivan, 2000). Indeed, both ToM and empathy rely on a related network of executive functions, including working memory (e.g. Morelli & Lieberman, 2012), and both have been shown to be automatically activated (to some degree) in response to social stimuli even in the absence of task-cues to keep track of others' mental states (e.g. Morelli & Lieberman, 2012; Schneider, Nott, & Dux, 2014). Second, neuroimaging research has revealed that making inferences about the mental and emotional states of story characters activates overlapping neuronal networks, including the medial prefrontal cortex, temporo-parietal junction and temporal poles (Völlm et al., 2006). Differences in brain activity between the two are thought to reflect the need to infer causality and intentions in ToM,

and emotional processing in empathy. Third, clinical groups who show impairments on ToM tasks (e.g. individuals with autism spectrum disorders, schizophrenia and psychopathy) also show a reduced ability to empathize, and lower scores on tests of empathy (e.g. Baron-Cohen & Wheelwright, 2004; Blair, 2005; Bora et al., 2008). Moreover, specific patterns of ToM/empathising deficits distinguish the different conditions, with impaired ToM more strongly related to autism, and impaired affective empathy predicting psychopathy (e.g. Jones et al., 2010; O'Nions et al., 2014). We aim to examine the degree to which empathy predicts the ability to process unfolding events according to other's beliefs in a healthy adult population.

While a great deal of research has been conducted to assess peoples' *explicit* responses to questions that probe understanding of false beliefs (e.g. Baron-Cohen et al., 1985; Birch & Bloom, 2007; Hogrefe et al., 1986; Wimmer & Perner, 1983), it is only recently that more sensitive *implicit* methods have enabled an exploration of the cognitive processes that underlie such decisions. For example, reaction time and eye-tracking techniques have demonstrated the speed with which belief inferences can be made, and have revealed the self/other biases that people display under different conditions (e.g. Apperly et al., 2006; Back & Apperly, 2010; Ferguson & Breheny, 2012; Ferguson et al., 2010; Kovács et al., 2010; Rubio-Fernandez & Glucksberg, 2011; Schneider et al., 2012). Further, a growing body of research has used electrophysiological methods (i.e. event related brain potentials, ERPs) to examine how false beliefs are reflected in the brain's electrical signal. The majority of these studies have examined the brain's response as participants answer explicit belief questions (e.g. "where does X think the Y is?"), and have demonstrated a frontally-distributed late slow wave (LSW) when people are required to reason about others' (false) beliefs versus reality (e.g. Liu et al., 2004; Sabbagh & Taylor, 2000; Wang et al., 2008; Zhang et al., 2009). This difference is thought to reflect the key processes that distinguish mental states from reality (Liu et al., 2004; Sabbagh & Taylor, 2000), including the experience of conflicting perspectives, and the need to inhibit the self-perspective when inferring others' beliefs (Zhang et al., 2009). More recent studies have examined ToM

under more implicit conditions- while participants observe pictorial sequences of events depicting beliefs and desires (Geangu et al., 2013; Kuhn-Popp et al. 2013; Meinhardt et al., 2012). Consistent with previous research, these passive studies have found a widely distributed LSW, which suggests that implicit monitoring of others' beliefs continues even in the absence of an explicit instruction to monitor mental states.

In contrast to these previous studies, the current study tested participants' understanding of false beliefs as they read narratives that described a story character having a true or false belief about the location of an object. Thus, we employed an anomaly detection reading paradigm while recording ERPs (i.e. the character looks for the object in the belief consistent or inconsistent location), which aimed to exploit the brain's clear sensitivity to stimulus predictability and semantic integration processes during language comprehension; the N400 effect (Kutas & Hillyard, 1980; Lau et al., 2013). This component is a centroparietally distributed, negative-going deflection in the ERP, which peaks approximately 400 ms after word-onset. Extensive research in psycholinguistics and cognitive neuroscience has shown that the amplitude of this ERP component is directly influenced by inconsistencies of both local and contextual information (e.g. Hagoort et al., 2004; Van Berkum et al., 2003). Moreover, typical N400 responses to local semantic anomalies (e.g. *the peanut was <u>in love</u>*) have been shown to reverse within an appropriate discourse context (Filik & Leuthold, 2013; Nieuwland & Van Berkum, 2006; Nieuwland, 2013; Nieuwland & Martin, 2012). Similar N400 effects are activated when a narrative describes a character's inappropriate emotional response to a given social situation (Leuthold, Filik, Murphy, & MacKenzie, 2012), or when s statement conflicts with a person's moral values (Van Berkum, Holleman, Nieuwland, Otten, & Murre, 2009). Here, we aim to establish for the first time whether the brain is sensitive to inconsistencies of other peoples' actions when they violate their beliefs, or whether the reader's own knowledge of events has a stronger influence on incremental processing. Thus, we compare N400 responses to belief-consistent and inconsistent events under true and false belief conditions, which is expected to

reveal listeners' *preferred* interpretations of the unfolding discourse in the earliest moments of processing. Importantly, this passive reading paradigm allowed us to examine the brain's immediate sensitivity to information that is consistent/inconsistent with beliefs, without making inferences about others' mental states an explicit part of the reader's task.

We recorded ERPs while participants read short narratives in which a character's (mistaken) belief conflicts with the participants' own knowledge about reality. An example of a *false* belief scenario is shown in (1) where reality and the beliefs of the story character are in direct conflict with one another.

(1) Gillian cooked a casserole and left it to cool down in the oven. While Gillian was not looking, Mark moved the casserole to the fridge. When Gillian wanted to eat the casserole, she looked in the fridge.

In this example, context suggests that Gillian won't know that the casserole has moved from the oven to the fridge (she was *not looking* while that happened), so her reported actions (s*he looked in the fridge*) are inconsistent with her beliefs. As described above, previous research has demonstrated that people are sensitive to others' beliefs even without being given an explicit instruction to track their mental states (e.g. Ferguson et al., under review; Ferguson & Breheny, 2012; Ferguson et al., 2010; Geangu et al., 2013; Schneider et al., 2012). Thus, if readers have already established a representation of the character's false belief based on the prior context, then some processing difficulty should be revealed when readers encounter the belief-inconsistent critical word (*fridge*). In line with the N400 literature described above, it is expected that such a difficulty will be reflected in an increased N400 effect following inconsistent words compared to consistent words (i.e. *oven* for the given example). However, if readers have not fully accommodated the character's false belief, they may process the incoming information egocentrically (i.e. biased to their own knowledge of reality), and instead show an increased N400 effect for the belief-consistent (but reality-inconsistent) word compared to the belief-inconsistent word. For comparison, 'true belief' passages where the character explicitly saw the

object get moved (e.g. 'Gillian spotted Mark move the casserole...') were included as a baseline of contextual anomaly detection, where we expect to see similar N400 effects for the inconsistent information based on readers' knowledge of narrative reality. Finally, if ability to integrate others' beliefs online is modulated by one's ability to empathise with others, then we expect to see a larger inconsistency detection response on false belief trials in individuals with high levels of empathy compared to those with low levels of empathy. Here, we use the Empathy Quotient questionnaire (Baron-Cohen & Wheelwright, 2004) as a measure of social aptitude, which indexes 'global empathy', including both cognitive empathy and emotional reactivity (Lawrence et al., 2004).

Method

Participants

Twenty-eight native English speakers from the University of Kent took part in this study (M_{age} = 20, SD_{age} = 3.9), and were either paid for participating or received course credits. Of these, 20 were female, and 25 were right-handed (handedness was measured using the Oldfield Edinburgh Handedness Inventory (Oldfield, 1971)). Participants did not have dyslexia and had vision that they reported to be normal or corrected-to-normal. All participants were naïve to the purpose of the study.

Materials and Design

One hundred and forty experimental items were created as in Table 1. Each item consisted of three sentences: Sentence one introduced a character and described that character putting a target object in a given location. Sentence two described a second character moving the target object to a new location. This action was either 'explicitly observed' or 'missed' by the first character, creating a true or false belief regarding the object's location for that character respectively (e.g. "Later, Janet saw Barry move the…", *versus* "While Janet was busy, Barry moved the…"). A

final third sentence described the first character looking for the object, and thus drew reference to a location that was either consistent or inconsistent with their true or false belief (e.g. "When Janet wanted to see the painting, she looked in the *kitchen/hall*"). This resulted in a 2 (belief: true *vs.* false) x 2 consistency: consistent *vs.* inconsistent) within subjects design. Note that reality-violating locations are congruent in a false belief context, and vice-versa.

Table 1: Example experimental item in each of the four conditions. Critical words are underlined for exposition only.

True Belief	**Consistent**	Janet unpacked the belongings and put the painting in the hall. Later, Janet saw Barry move the painting to the kitchen. When Janet wanted to see the painting, she looked in the kitchen.
	Inconsistent	Janet unpacked the belongings and put the painting in the hall. Later, Janet saw Barry move the painting to the kitchen. When Janet wanted to see the painting, she looked in the hall.
False Belief	**Consistent**	Janet unpacked the belongings and put the painting in the hall. While Janet was busy, Barry moved the painting to the kitchen. When Janet wanted to see the painting, she looked in the hall.
	Inconsistent	Janet unpacked the belongings and put the painting in the hall. While Janet was busy, Barry moved the painting to the kitchen. When Janet wanted to see the painting, she looked in the kitchen.

Experimental items were tested using a pre-test for cloze probability using an online questionnaire platform (Qualtrics). This allowed us to ensure that adult participants would correctly predict the belief appropriate location for both true belief and false belief conditions, and to test whether offline differences in this ability existed. Twenty-two students from the University of Kent completed the pre-test, which consisted of ten passages depicting a character with a true belief, and ten passages depicting a character with a false belief (two counterbalanced lists meant that eleven participants completed each list, with one version of each item appearing in each list). Items were presented one at a time, truncated before the final critical word, and participants were instructed to complete the sentence with the first sensible word coming to mind.

Cloze probability was computed as the percentage of trials that elicited the intended consistent or inconsistent critical words. Mean cloze probability scores for the consistent word in each condition revealed high accuracy and no significant difference between true belief (M = .96, SD = .07) and false belief (M = .94, SD = 13; $t(21)$ = .64, p = .53) contexts. In addition, low-level properties of the sentence-final critical words were matched according to word length, log-frequency, and familiarity (using the MRC Psycholinguistics Database; Wilson, 1988). Statistical comparisons between conditions found no significant differences in any of these measures (All ts < 1; see Table 2 for mean values on each measure).

Table 2: Mean pre-test ratings per condition for the final set of experimental items. Standard deviations are shown in parentheses.

	True Belief		False Belief	
	Consistent	Inconsistent	Consistent	Inconsistent
Cloze Probability	0.96 (.08)	0.04 (.08)	0.94 (.13)	0.06 (.13)
Word length	5.9 (2.0)	6 (1.7)	6 (1.7)	5.9 (2.0)
Word Frequency (log)	2.89 (1.5)	3.22 (1.6)	3.22 (1.6)	2.89 (1.5)
Word Familiarity	549.7 (57.6)	560.7 (54.0)	560.7 (54.0)	549.7 (57.6)

Four presentation lists were then created, with each list containing one hundred and forty experimental items, thirty-five in each of the four conditions. The one hundred and forty experimental items in each list were interspersed randomly among sixty-eight unrelated filler sentences to create a single random order and each subject only saw each target sentence once, in one of the four conditions. Seven participants were randomly assigned to read each list.

All participants also completed the Empathy Quotient questionnaire (Baron-Cohen & Wheelwright, 2004), as a measure of social aptitude. The empathy questionnaire contains 40 statements (e.g. “I can easily tell if someone else wants to enter a conversation”), and participants indicated the degree to which each statement relates to them (on a 4-point scale: “strongly agree”,

"slightly agree", "slightly disagree" and "strongly disagree". Each participant received a score on a scale of 0 to 80, using a scoring key designed by Baron-Cohen and Wheelwright (2004), where a low score indicates low levels of empathy and a high score indicates high levels of empathy. Participant scores in the current sample averaged 44.4 and ranged from 21 to 68.

Procedure

Participants were informed about the EEG procedure and experimental task. After electrode application they were seated in a booth where they read the materials from a computer screen. There were four practice trials to familiarize them with the procedure, after which the experimenter answered any questions. Each trial began with the presentation of a single centrally-located red fixation cross for 500ms to signal the start of a new trial. After this time, a white fixation cross appeared for 500ms. Next, the first two context sentences were presented on the screen, and participants were instructed to read these sentences and press spacebar on a keyboard to continue when ready. A blank screen appeared for 500ms, followed by a fixation cross (500ms). The third target sentence was then presented word-by-word, with each word appearing at the centre of the screen for 300ms, with a 200ms blank-screen interval between words. Target words were always sentence final, and thus appeared with a full stop. A 2500ms blank-screen interval followed each item. As recommended by Van Berkum (2004, 2012), there was no secondary task to verify attention, since secondary tasks have the potential to recruit their own brain responses that might interfere with the brain activity under examination. Trials appeared in eight blocks of twenty-six trials. Each block was separated by a break, the duration of which was determined by the participant. At the end of the main EEG task, participants completed the Empathy Quotient questionnaire. Thus, participants were tested in a single session that lasted approximately one hour, during which they were seated in a comfortable chair located in an isolated room.

Electrophysiological Measures

A Brain Vision Quickamp amplifier system was used with an ActiCap cap for continuous recording of electroencephalographic (EEG) activity from 62 active electrodes over midline electrodes Fz, Cz, CPz, Pz, POz, and Oz, over the left hemisphere from electrodes Fp1, AF3, AF7, F1, F3, F5, F7, FC1, FC3, FC5, FC7, C1 C3, C5, T7, CP1, CP3, CP5, TP7, A1, P1, P3, P5, P7, PO3, PO7, PO9, O1, and from the homologue electrodes over the right hemisphere. EEG and EOG recordings were sampled at 1000 Hz, and electrode impedance was kept below 10kΩ. Off-line, all EEG channels were recalculated to an average mastoid reference.

Prior to segmentation, EEG and EOG activity was band-pass filtered (0.01-30 Hz, 12 dB/oct), and EEG activity containing blinks was corrected using a semi-automatic ocular ICA correction approach (Brain Vision Analyzer 2). The continuous EEG record was then segmented into epochs of 1200ms, starting 200ms prior to the onset of the target word. Thus, the post-stimulus epoch lasted for a total duration of 1000ms. Semi-automatic artifact detection software (Brain Vision Analyzer 2) was run, to identify and discard trials with non-ocular artifacts (drifts, channel blockings, EEG activity exceeding ± 75μV). This procedure resulted in an average trial-loss of 6.5% per condition.

ERP Data Analysis

For analysis of the EEG data, the signal at each electrode site was averaged separately for each experimental condition time-locked to the onset of the target word. Before the measurement of ERP parameters, the waveforms were aligned to a 200ms baseline prior to the onset of the target word. To analyze experimental effects on the N400, mean ERP amplitude was determined in the time interval from 250-400ms relative to target word onset.

ERP amplitudes over lateral electrodes were analysed using four regions of interest (ROIs). Given the broad distribution of the N400, and in line with recent analyses of narrative comprehension (e.g. Nieuwland, 2013), electrodes were divided along a left-right dimension, and

an anterior-posterior dimension. The two ROIs over the left hemisphere were: left-anterior (Fp1, AF3, AF7, F1, F3, F5, F7, FC1, FC3, FC5, FT7), and left-posterior (CP1, CP3, CP5, TP7, P1, P3, P5, P7, PO3, PO7, O1); two homologue ROIs were defined for the right hemisphere. ERP amplitudes over midline electrodes (Fz, Cz, CPz, Pz, POz, Oz), where the N400 is maximal, were analysed separately from data recorded over lateral electrode sites.

For the statistical analysis of the N400 in each condition, an ANOVA was performed over lateral electrodes with variables belief (true *vs.* false), consistency (consistent *vs.* inconsistent), hemisphere (left *vs.* right), and ant-pos (anterior *vs.* posterior). ERP amplitudes over midline electrodes were analysed using a belief (true *vs.* false) x consistency (consistent *vs.* inconsistent) x electrode (Fz, Cz, CPz, Pz, POz, Oz) ANOVA. To examine the effect of individual differences in empathy on belief understanding, Pearson's correlations were performed comparing participants' empathy scores with the 'inconsistency effect' for true belief and false belief conditions separately. The 'inconsistency effect' was calculated by subtracting the N400 amplitude for the consistent condition from the N400 amplitude for the inconsistent condition between 250-400ms post-target word onset, over all electrode sites. Thus, a negative score indicates a larger N400 effect for the inconsistent compared to consistent condition (i.e. appropriate anomaly detection), and a positive score indicates a large effect for the consistent compared to inconsistent condition (i.e. interpreting events egocentrically in false belief contexts).

Results

N400 effect Analyses

Grand average ERP waveforms over three midline electrodes (Fz, Cz and Pz) are presented in Figure 1. It can be seen that following a true belief context, inconsistent target words triggered a more negative-going deflection (N400) than consistent target words, starting between 200-250ms after critical word onset. In contrast, this pattern appears to be reversed for the false belief

condition, with consistent target words eliciting a slightly more negative-going N400 component within the same time window following target word onset.

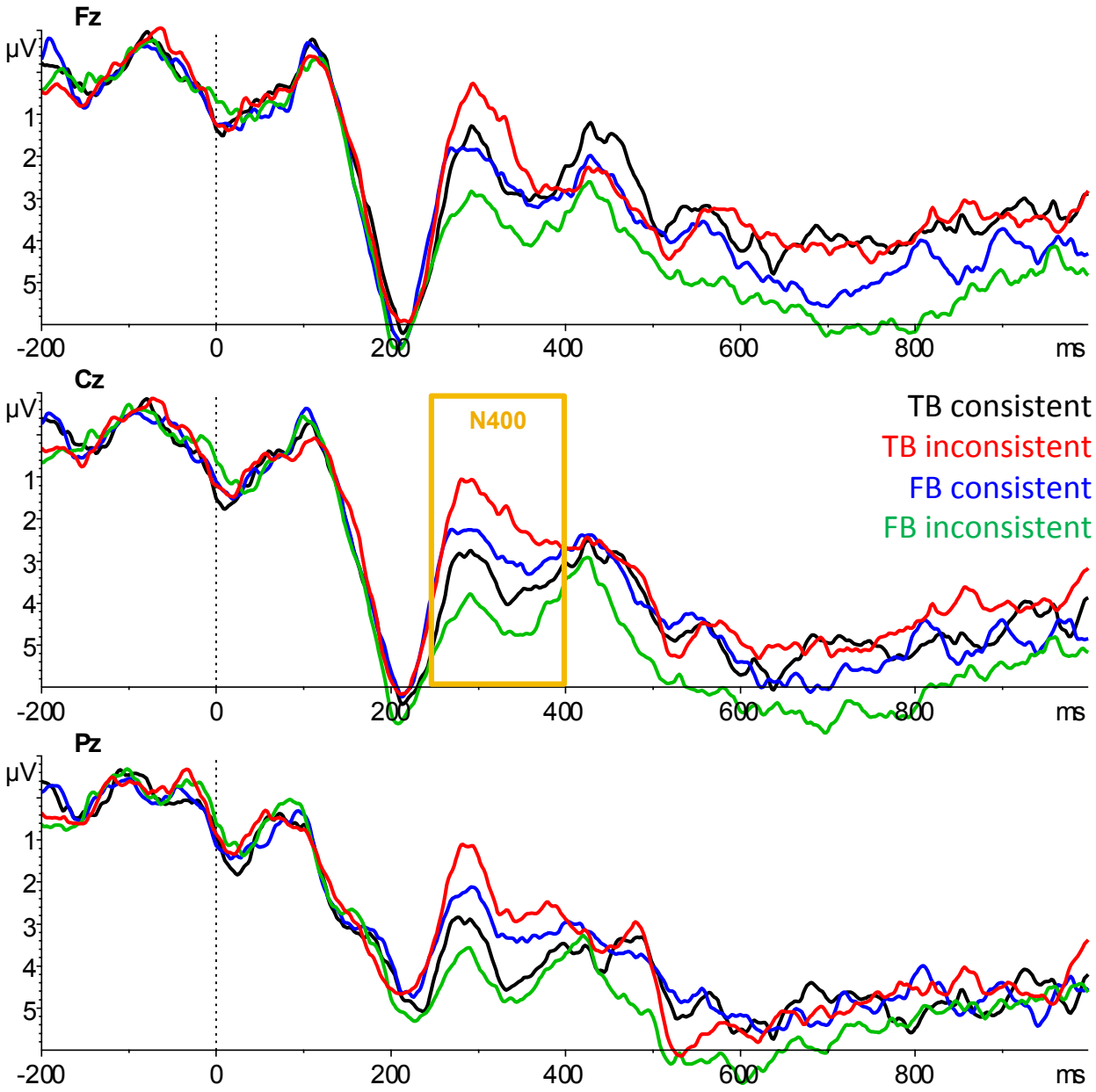

Figure 1: Grand average ERPs over midline electrodes elicited by critical words in the target sentence for each of the four conditions. Note that negativity is plotted upwards.

Analysis of the N400 amplitude over lateral electrodes in the 250-400ms time interval revealed a main effect of belief [$F(1, 27) = 5.75$, $p < .03$, ${}_p\eta^2 = .18$], such that overall, ERP

waveforms were more negative-going for true belief contexts compared to false belief context (2.29 *vs*. 2.82 μV). Moreover, there was a belief * consistency interaction [$F(1, 27) = 13.68$, $p < .001$, $_p\eta^2 = .34$], which also appeared in a three-way interaction with ant-pos [$F(1, 27) = 7.49$, $p < .01$, $_p\eta^2 = .22$]. To examine these effects further we conducted simple main effects analyses to compare the consistency effects at each context level, separately for anterior and posterior electrode sites. Results at anterior sites revealed no significant difference between true belief consistent and inconsistent conditions ($t(27) = 1.04$, $p = .31$), or between false belief consistent and inconsistent conditions ($t(27) = -1.97$, $p = .06$). In contrast, at posterior sites (where the N400 is typically maximal), a clear effect of consistency emerged for both true belief ($t(27) = 4.36$, $p < .001$) and false belief ($t(27) = -2.51$, $p < .02$) context conditions. In the true belief condition, this reflected the expected increased N400 amplitude following a belief-inconsistent target word compared to a belief-consistent target word (2.06 *vs*. 3.38 μV). However, in the false belief condition, the N400 amplitude was largest following a belief-consistent target word compared to a belief-inconsistent target word (2.77 *vs*. 3.65 μV).

Over midline electrodes, the main effect of belief was again significant [$F(1, 27) = 5.22$, $p < .03$, $_p\eta^2 = .16$], reflecting a more negative-going ERP waveform for true belief contexts compared to false belief contexts (2.57 *vs*. 3.24 μV). In addition, the interaction between belief and consistency was significant [$F(1, 27) = 15.84$, $p < .001$, $_p\eta^2 = .37$]. Simple main effects analyses compared the consistency effects at each context level, and revealed that while the true belief context elicited a larger N400 for inconsistent versus consistent target words ($t(27) = 3.33$, $p < .003$; 1.93 *vs*. 3.21 μV), the false belief context elicited the reverse pattern, with a larger N400 for consistent versus inconsistent target words ($t(27) = -3.05$, $p < .005$; 2.61 *vs*. 3.87 μV). Scalp topographies of the consistency effect in each condition are shown in Figure 2.

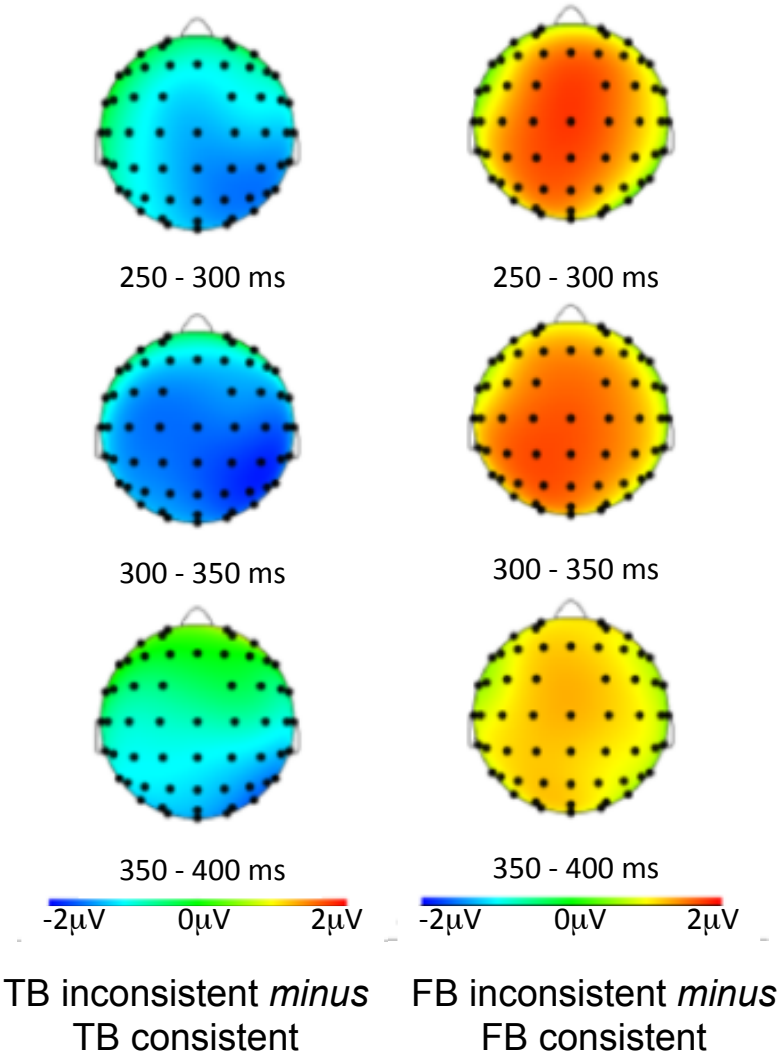

Figure 2: Topographic maps of the ERP difference waveform for each belief context condition (inconsistent *minus* consistent) between 250-400ms from critical word onset.

Correlations with empathy

Correlations compared participants' empathy scores with the amplitude and valence of the 'inconsistency effect' for true belief and false belief conditions separately. This data can be seen in Figure 3. Recall that a negative N400 effect typically reflects lower expectancy, or difficulty integrating a word into the wider context, thus a negative score indicates a larger N400 effect for the inconsistent compared to consistent condition (i.e. appropriate anomaly detection), and a positive score indicates a large effect for the consistent compared to inconsistent condition (i.e.

interpreting events egocentrically in false belief contexts). In the true belief condition, there was no correlation between empathy and the inconsistency effect ($r(26) = .03$, $p = .43$), demonstrating that empathy does not predict the detection of inconsistencies within a true context. In contrast, a significant negative relationship between empathy score and direction of the N400 inconsistency effect was found in the false belief condition ($r(26) = -.51$, $p < .005$). This suggests that individuals with lower empathy scores interpret unfolding events egocentrically, but individuals with higher empathy scores are more likely to successfully interpret events according to the character's beliefs (thus showing an appropriately larger N400 response to the false belief inconsistent condition compared to the consistent condition), possibly alongside an interpretation of events according to reality. A comparison of the slopes for true and false belief conditions (Steiger, 1980) revealed a marginal difference between the correlation slopes ($z = 1.6$, $p = .11$). Considered within the context of a significant empathy-inconsistency correlation effect for false belief but not true belief contexts, this shows that while the slopes are in the same direction, there is some difference in the magnitude of the relationship when directly compared.

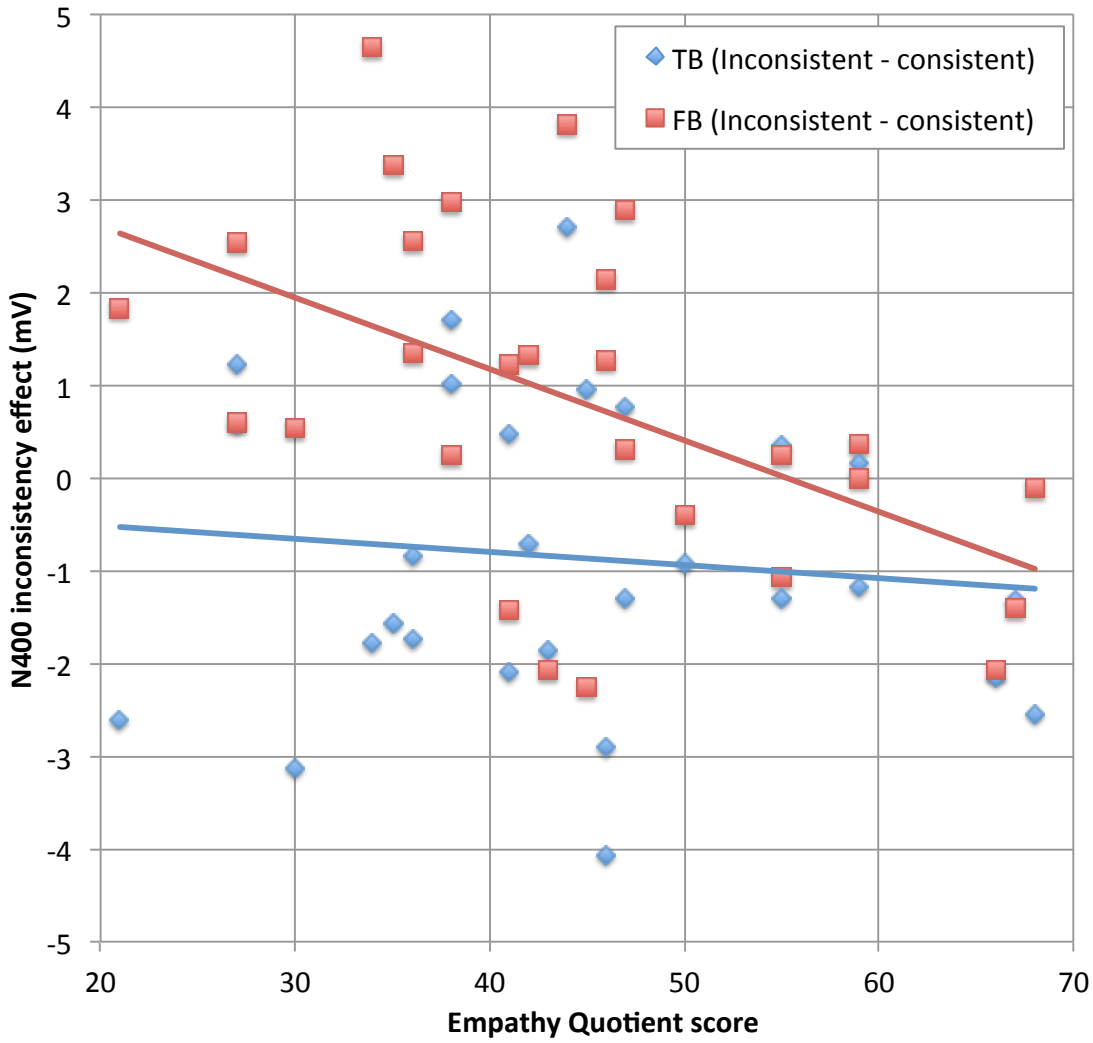

Figure 3: Correlation between individuals' empathy quotient scores and the N400 inconsistency effect for true and false belief conditions. The N400 inconsistency effect is calculated as the difference in N400 amplitude (inconsistent *minus* consistent) between 250-400ms post-target word onset, over all electrode sites.

Discussion

Understanding others' beliefs frequently requires the comprehender to represent a version of the world that is inconsistent with their own knowledge of reality. However, rapid dissociation of these two types of information is important for successful everyday communication, otherwise our own knowledge of reality would become confounded with others peoples' beliefs. In this

paper, we report the results of an ERP experiment whereby participants read narratives that described a story character having a true or false belief about the location of an object, and manipulated the consistency of their actions based on this belief. By comparing N400 responses to belief-consistent and inconsistent events under true and false belief conditions, we aimed to investigate for the first time whether the brain's response is sensitive to violations of another person's mind, or whether the reader's own knowledge of reality has a stronger influence on incremental processing. Subsequent analyses examined the degree to which individual differences among participants influence this response by assessing how one's ability to empathise with others modulates the ability to integrate others' beliefs online.

Results showed that when the character held a true belief about the object's location, the N400 was more negative-going for belief inconsistent *vs.* belief consistent location nouns. This finding fits with previous research, which has shown that readers are sensitive to discourse-level inconsistencies, even when that information fits the sentence on a local level (i.e. the local sentence is grammatical and semantically correct; e.g. Camblin, Gordon, & Swaab, 2007; Van Berkum, Zwitserlood, Hagoort, & Brown, 2003). This demonstrates that our readers were correctly considering the wider discourse context at the point that they integrated the character's actions. However, note that in this true belief condition, the reader's and the character's beliefs were aligned with reality, thus no interference was present to disrupt processing (except possibly a memory trace of the object in its initial location). Moreover, readers could correctly integrate the character's actions based on their own knowledge about reality, without consideration of the character's beliefs. Therefore, in order to test readers' ability to interpret unfolding events according to the character's beliefs, we must focus on their responses when the character held a false belief about the object's location.

The reverse pattern of effects was found when the character held a false belief about the object's location; the N400 was more negative-going for belief consistent *vs.* belief inconsistent location nouns. This suggests that readers do not immediately integrate the character's beliefs,

but instead initially rely on their own egocentric knowledge of the object's true location when processing described events. This finding fits with previous suggestions of an initial egocentric bias or 'pull of reality' in ToM use (Birch & Bloom, 2004, 2007; Ferguson & Breheny, 2012; Keysar, Barr, Balin, & Brauner, 2000; Mitchell, Robinson, Isaacs, & Nye, 1996), whereby knowledge of the object's actual location delayed readers' access to the belief inference. Nevertheless, in a separate offline sentence completion task, participants were 94% accurate in stating the correct false belief-appropriate location, showing no difference in accuracy compared to the true belief condition. This suggests that readers can make the appropriate inference about a character's actions based on their beliefs when sufficient time is available, despite the initial interpretation of events relying on one's own knowledge.

However, further analyses revealed that individual differences in empathy influenced the magnitude, and to some degree the direction, of this inconsistency effect in the false belief condition (but not the true belief condition), therefore showing differences in the perspective preferences that readers adopt while interpreting narratives about (false) beliefs. Here, correlation analyses revealed that individuals with high levels of empathy were able to rapidly integrate contextual information about the character's beliefs and subsequently processed incoming information in terms of that belief (i.e. showed a more negative N400 for false belief inconsistent *vs.* consistent location nouns) alongside their own inference based on knowledge of reality (as reflected in inconsistency effects around zero). In contrast, individuals with low levels of empathy predominantly interpreted events in terms of their real world knowledge. Note that due to the relatively small sample size used here (N=28), statistical comparison between low and high empathisers (using a median split) was not possible. The lack of a correlation in the true belief condition demonstrates that empathy does not simply enhance one's general language comprehension skills, but that it relates specifically to one's ability to infer and use ToM online.

This study is the first to show that empathy is related to the degree to which people experience intrusions from their own knowledge/reality online. Such a relationship makes sense

given that both false belief reasoning and empathy recruit related processes of perspective-taking (i.e. understanding and predicting events in terms of other peoples' mental states- including their knowledge or emotional state). Indeed, both the ability to infer false beliefs and the ability to empathise with others recruit overlapping executive skills, including inhibition (of one's own mental state) and working memory (to represent the multiple mental states). This relationship between empathy and beliefs demonstrates that an egocentric or reality bias is not a default process in ToM use, and that such biases can be over-ridden when other peoples' perspectives are more appropriate for understanding and the comprehender possesses sufficient social and cognitive skills to inhibit this bias (see Brown-Schmidt, 2009; Brunye et al., 2012; Cane, Ferguson, & Apperly, under review; German & Hehman, 2006; Kessler & Wang, 2012; Lin et al., 2010). Whilst these findings show clear evidence of a relationship between empathy and the use of ToM online, we cannot assume a causal role; it is equally plausible that increased ToM use leads to increased empathy or that greater empathy leads to increased ToM use. Further research is needed to establish the existence and nature of such a causal relationship between empathy and ToM, and to understand the mechanisms that might underlie this relationship.

Taken together, these results demonstrate that when an individual's social skills are high the brain can be immediately sensitive to violations of other peoples' mental states, specifically their beliefs, and this can modulate the amplitude of the N400 effect, which is typically associated with stimulus predictability and semantic integration processes during language comprehension. It is interesting to note that this modulation occurred even in a passive reading task such as this where ToM use was not an explicit part of the task, and in fact did not benefit participants (c.f. Liu et al., 2004; Sabbagh & Taylor, 2000; Wang, Liu et al., 2008; Zhang et al., 2009). Although modulating a different ERP component (due to task differences), this finding fits with recent studies that have examined ToM under implicit conditions, and have reported evidence of implicit monitoring of others' beliefs without explicit instructions to track others' mental states (Geangu et al., 2013; Kuhn-Popp et al., 2013; Meinhardt et al., 2012).

In conclusion, when a character's described actions are inconsistent with respect to their true belief about reality this rapidly elicits processing difficulties during reading, as revealed by an enhanced N400 anomaly detection brain response. More interesting is the finding that when readers experience a conflict between their own knowledge of reality and a character's false belief, processing can be biased towards either the reality or belief-appropriate interpretation, depending on an aspect of individuals' social competence (i.e. empathy). Specifically, high empathisers successfully interpreted events according to the character's false beliefs (possibly alongside their own knowledge of reality), but low empathisers relied on their egocentric knowledge and therefore did not initially use the character's belief to interpret unfolding events. This study demonstrates the benefits of employing implicit ToM tasks and online measures in healthy adult populations, and brings to the broad field of ToM a new reading paradigm for studying mutual knowledge/ common ground phenomena. Finally, this study is the first to demonstrate the effect of empathy in adult online false belief understanding, and therefore illustrates the importance of considering individual differences when assessing ToM.

References

Apperly, I.A., Riggs, K.J., Simpson, A., Samson, D., & Chiavarino, C. (2006). Is belief reasoning automatic? *Psychological Science, 17*, 841-844.

Back, E. & Apperly, I. (2010). Two sources of evidence on the non-automaticity of true and false belief ascription. *Cognition, 115*, 54-70.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition, 21*, 37-46.

Baron-Cohen, S. & Wheelwright, S. (2004). The empathy quotient: an investigation of adults with Asperger syndrome or high functioning autism and normal sex differences. *Journal of Autism and Developmental Disorders, 34*, 163-175.

Birch, S.A.J. & Bloom, P. (2004). Understanding children's and adults' limitations in reasoning

about the mind. *Trends in Cognitive Sciences, 8*, 255-260.

Birch, S.A.J. & Bloom, P. (2007). The Curse of Knowledge in Reasoning About False Beliefs. *Psychological Science*, *18*, 382-386.

Blair, R.J.R. (2005). Responding the emotions of the others: dissociating the forms of empathy through the study of typical and psychiatric populations. *Consciousness and Cognition, 14*, 698-718.

Bora, E., Gökçen, S., & Veznedaroglu, B. (2008). Empathic abilities in people with schizophrenia. *Psychiatry Research, 160*, 23-29.

Brown-Schmidt, S. (2009). The role of executive function in perspective-taking during on-line language comprehension. *Psychonomic Bulletin and Review, 16*, 893-900.

Brunye T.T., Ditman T., Giles G.E., Mahoney C.R., Kessler K., & Taylor H.A. (2012). Gender and autistic personality traits predict perspective-taking ability in typical adults. *Personality and Individual Differences, 52*, 84-88.

Camblin, C.C., Gordon, P.C., & Swaab, T.Y. (2007). The interplay of discourse congruence and lexical association during sentence processing: Evidence from ERPs and eye tracking. *Journal of Memory & Language, 56*, 103-128**.**

Cane, J.E., Ferguson, H.J., & Apperly, I. (under review). Examining the influence of working memory and motivation on the time course of eye-movements in a perspective taking task. *Journal of Memory & Language*.

Converse, B. A., Lin, S., Keysar, B., & Epley, N. (2008). In the mood to get over yourself: mood affects theory-of-mind use. *Emotion*, *8*, 725-730.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*(1), 113–126. doi:10.1037/0022-3514.44.1.113

de Villiers, J. G. & Pyers, J. E. (2002). Complements to cognition: A longitudinal study of the relationship between complex syntax and false-belief-understanding. *Cognitive Development, 17*, 1037-1060.

Dymond, R. F. (1949). A scale for the measurement of empathic ability. *Journal of Consulting Psychology*, *13*, 127-133.

Ferguson, H.J., Apperly, I., Ahmad, J., Bindemann, M., & Cane, J.E. (under review). Task constraints distinguish perspective inferences from perspective use during discourse interpretation in a false belief task. *Cognition.*

Ferguson, H.J. & Breheny, R. (2012). Listeners' eyes reveal spontaneous sensitivity to others' perspectives. *Journal of Experimental Social Psychology*, *48*, 257-263.

Ferguson, H.J., Scheepers, C., & Sanford, A.J. (2010). Expectations in counterfactual and theory of mind reasoning. *Language and Cognitive Processes, 25*, 297-346.

Filik, R. & Leuthold, H. (2013). The role of character-based knowledge in online narrative comprehension: Evidence from eye movements and ERPs. *Brain Research, 1506*, 94-104.

Flavell, J.H., Flavell, E.R., Green, F.L., & Moses, L.J. (1990). Young children's understanding of fact beliefs versus value beliefs. *Child Development, 61*, 915-928.

Geangu, E., Gibson, A., Kaduk, K., & Reid, V.M. (2013). The neural correlates of passively viewed sequences of true and false beliefs. *Social Cognitive and Affective Neuroscience, 8*, 432-437.

German, T. & Hehman, J.A. (2006). Representational and executive selection resources in 'theory of mind': Evidence from compromised belief-desire reasoning in old age. *Cognition*, *101*, 129-152.

Hagoort, P., Hald, L., Bastiaansen, M.C.M., & Petersson, K.M. (2004). Integration of Word Meaning and World Knowledge in Language Comprehension. *Science, 304,* 438-440.

Hogrefe G.J., Wimmer H., & Perner J. (1986). Ignorance versus false belief: A developmental lag in attribution of epistemic states. *Child Development*, 57, 567-582.

Jones, A.P., Happe, F.G.E., Gilbert, F., Burnett, S., Viding, E. (2010). Feeling, caring, knowing: different types of empathy deficit in boys with psychopathic tendencies and autism spectrum disorder. *Journal of Child Psychology and Psychiatry, 51*, 1188-1197.

Kessler, K. & Wang, H. (2012). Spatial perspective taking is an embodied process, but not for everyone in the same way: differences for gender and social skills score. *Spatial Cognition and Computation, 12*, 133-158.

Keysar, B., Barr, D.J., Balin, J.A., & Brauner, J.S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science, 11*, 32-38.

Kovács, Á.M., Téglás, E., & Endress, A.D. (2010). The social sense: susceptibility to others' beliefs in human infants and adults. *Science, 330*, 1830-1834.

Kühn-Popp, N., Sodian, B., Sommer, M., Döhnel, K., & Meinhardt, J. (2013). Same or different? ERP correlates of pretense and false belief reasoning in children. *Neuroscience, 248*, 488-498.

Kutas, M. & Hillyard, S.A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science, 207*, 203-205.

Lau, E.F., Holcomb, P.J., & Kuperberg, G.R. (2013). Dissociating N400 effects of prediction from association in single-word contexts. *Journal of Cognitive Neuroscience, 25*, 484-502.

Lawrence, E.J., Shaw, P., Baker, D., Baron-Cohen, S., & David, A.S. (2004). Measuring Empathy - reliability and validity of the empathy quotient. *Psychological Medicine, 34*, 911-919.

Leuthold, H., Filik, R., Murphy, K., & Mackenzie, I.G. (2012). The on-line processing of socio-emotional information: Inferences from brain potentials. *Social Cognitive and Affective Neuroscience, 7,* 457-466.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology, 46*, 551-556.

Liu, D., Sabbagh, M.A., Gehring, W.J., & Wellman, H.M. (2004). Decoupling beliefs from reality in the brain: an ERP study of theory of mind. *NeuroReport, 15*, 991-995.

Meinhardt, J., Kühn-Popp, N., Sommer, M., & Sodian, B. (2012). Distinct neural correlates underlying pretense and false belief reasoning: evidence from ERPs. *NeuroImage, 63*, 623-631.

Mitchell, P., Robinson, E.J., Isaacs, J.E., & Nye, R.M. (1996). Contamination in reasoning about false belief: an instance of realist bias in adults but not children. *Cognition*, *59*, 1-21.

Morelli, S.A. & Lieberman, M.D. (2013). The role of automaticity and attention in neural processes underlying empathy for happiness, sadness, and anxiety. *Frontiers in Human Neuroscience, 7*, 1.

Nieuwland, M.S. (2013). "If a lion could speak …": Online sensitivity to propositional truth-value of unrealistic counterfactual sentences. *Journal of Memory & Language, 68*, 54-67.

Nieuwland, M.S. & Martin, A.E. (2012). If the real world were irrelevant, so to speak: The role of propositional truth-value in counterfactual sentence comprehension. *Cognition, 122*, 102-109.

Nieuwland, M.S. & Van Berkum, J.J.A. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, *18,* 1098-1111.

O'Nions, E., Sebastian, C.L., McCrory, E., Chantiluke, K., Happé, F., & Viding, E. (2014). Neural bases of Theory of Mind in children with autism spectrum disorders and children with conduct problems and callous-unemotional traits. *Developmental Science*.

Rubio-Fernández, P. & Glucksberg, S. (2011). Reasoning about other people's beliefs: Bilinguals have an advantage..*Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 211-217.

Sabbagh, M.A. & Taylor, M. (2000). Neural correlates of theory-of-mind reasoning: an event-related potential study. *Psychological Science, 11*, 46-50.

Savitsky, K., Keysar, B., Epley, N., Carter, T., & Swanson, A. (2011). The closeness-communication bias: Increased egocentrism among friends versus strangers. *Journal of Experimental Social Psychology*, *47*, 269-273.

Schneider, D., Bayliss, A.P., Becker, S., & Dux, P.E. (2012). Sustained implicit belief processing revealed by eye movements. *Journal of Experimental Psychology: General, 141*, 433-438.

Steiger, J.H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245-251.

Stotland, E. (1969). Exploratory investigations of empathy. In L.Berkowitz (Ed.), *Advances in Experimental Social Psychology* (*Vol. 4*, pp. 271– 312). New York: Academic Press.

Tager-Flusberg, H. & Sullivan. K. (2000). A componential view of theory of mind: evidence from syndrome. *Cognition, 76*, 59-90.

Van Berkum, J.J.A. (2012). The electrophysiology of discourse and conversation. In M. J. Spivey, K. McRae, & M. F. Joanisse (Eds.), *The Cambridge handbook of psycholinguistics* (pp. 589-614). New York: Cambridge University Press.

Van Berkum, J. J. A. (2004). Sentence comprehension in a wider discourse: Can we use ERPs to keep track of things? In M. Carreiras, Jr., & C. Clifton (Eds.), *The on-line study of sentence comprehension: eyetracking, ERPs and beyond* (pp. 229-270). New York: Psychology Press.

Van Berkum, J. J. A., Holleman, B., Nieuwland, M. S., Otten, M., & Murre, J. (2009). Right or wrong? The brain's fast response to morally objectionable statements. *Psychological Science, 20*, 1092-1099.

Van Berkum, J.J.A., Zwitserlood, P., Hagoort, P., & Brown, C.M. (2003). When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect. *Cognitive Brain Research, 17,* 701-718.

Völlm, B.A., Taylor, A.N., Richardson, P., Corcoran, R., Stirling, J., McKie, S., Deakin, J.F.W., & Elliott, R. (2006). Neuronal correlates of theory of mind and empathy: a functional magnetic resonance imaging study in a nonverbal task. *Neuroimage, 29*, 90-98.

Wang, Y.W., Liu, Y., Gao, Y.X., Chen, J., Zhang, W.X., & Lin, C.D. (2008). False belief reasoning in the brain: An ERP study. *Science China Life Sciences, 51*, 72-79.

Wellman, H.M. & Bartsch, K. (1988). Young children's reasoning about beliefs. *Cognition, 30*, 239-277.

Wellman, H.M., Cross, D., & Watson, J. (2001). Meta-Analysis of Theory-of-Mind Development: The Truth about False Belief. *Child Development*, *72*, 655-684.

Wimmer, H. & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*, 103-128.

Wu, S. & Keysar, B. (2007). The Effect of Culture on Perspective Taking. *Psychological Science*, *18*, 600-606.

Zhang, T., Sha, W., Zheng, X., Ouyang, H., & Li, H. (2009). Inhibiting one's own knowledge in false belief reasoning: An ERP study. *Neuroscience Letters, 467*, 194-198.